Topic Tracking for Radio, TV Broadcast, and Newswire

Hubert Jin, Rich Schwartz, Sreenivasa Sista, Frederick Walls

BBN Technologies
70 Fawcett Street, Cambridge, MA 02138
hjin@bbn.com

ABSTRACT

We present our tracking system for the 1998 Topic Detection and Tracking project (TDT-2). This project addresses multiple sources of information in the form of both text and speech from newswire, radio and television news broadcast programs. The technical challenge of TDT-2 tracking is to follow the topics being discussed in the stories from multiple sources. Our tracking system is probability based and we successfully solve the problem of score normalization across topics. Our automatic score normalization is simple, efficient and very effective. Tested on the 20K TDT-2 stories collected between March and April 1998, our tracking system achieves the performance of 1.5% miss error (on a combination of closed caption and newswire) and 3.0% miss error (on a combination of automatic speech recognition output and newswire) at the cost of 0.1% false alarm error. In the 1998 TDT-2 evaluation, our tracking system was ranked the best with the official topic-weighted Ctrack of 0.0057.

1. INTRODUCTION

The TDT program deals with stories in the form of both text and speech from newswire, radio and television news broadcast programs. In 1997, the TDT Pilot Project (also called TDT-1) only addressed newswire text. Quite a lot of research has been done in this new IR area [5]. Starting in 1998, speech data has been added in the corpus. There are about 60K stories in the TDT-2 corpus, collected from January to June, 1998. Roughly two thirds of the stories are in both forms of speech and text. The TDT-2 corpus is partitioned into 3 parts; each has a span of two months. Our systems were mostly developed and tested on the dev-test data, i.e. the 20K stories from March to April. The 1998 TDT-2 evaluation was performed on the data from the last two months. LDC (http://www.ldc.upenn.edu) annotated the corpus, and Dragon Systems (http://www.dragonsys.com) did automatic speech recognition for all the speech data [4]. The average word error from Dragon's automatic speech recognition was about 23%.

The purpose of the 1998 Topic Detection and Tracking project is to advance the state of the art in technologies required to segment, detect, and track topical information in an information stream, so that old topics can be tracked and new ones be detected on a variety of media sources including radio, TV broadcast, and Newswire [2] [3] [1]. The general TDT task domain is to be explored and technology is to be developed in the context of an evaluation-driven R&D paradigm, in which key technical challenges are defined and supported by formal evaluations. Three key technical challenges: Topic Segmentation, Topic Detection, and Topic Tracking, are explored in TDT-2. This paper deals with the topic tracking task.

In tracking, the system is given a few (4 is used as the default evaluation condition) stories about a particular topic, and also lots of other stories which are known to be irrelevant to the topic. The goal is to produce a score for each remaining story in the corpus that indicates how likely it is to be on the target topic, as well as a decision whether this story is about on the topic.

The paper is organized into three main sections. First of all, we introduce the architecture of the 1998 TDT tracking system. In the following section, we discuss the probabilistic framework for comparing one news story with a group of news stories. The third section contains the specific techniques we used in the TDT tracking project. Finally, we present our research and evaluation results in the third section.

2. System Description

We developed various individual systems for topic tracking based on different probabilistic models, each likely to focus on a different aspect of the underlying truth. We believe that an optimal system should include many probabilistic measures, combined in an appropriate way. We have developed three individual tracking systems - Topic Spotting (TS), Information Retrieval (IR) and Relevance Feedback (RF). Unsupervised adaptation has been implemented for TS and

RF. Scores from each individual system are self normalized, within each topic, using the labeled off-topic stories. We use logistic regression modeling to estimate combination coefficients. Figure 1 illustrates the architecture of the 1998 BBN TDT tracking system:

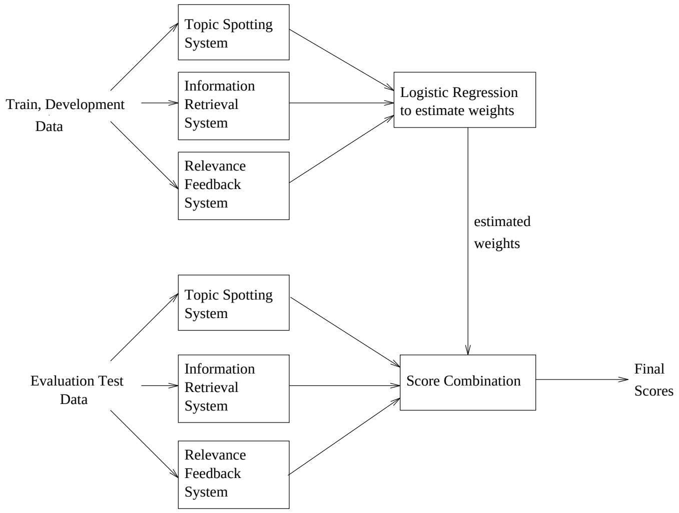
Train, Development Data
Topic Spotting System
Information Retrieval System
Relevance Feedback System
Logistic Regression to estimate weights
estimated weights
Evaluation Test Data
Topic Spotting System
Information Retrieval System
Relevance Feedback System
Score Combination
Final Scores

Figure 1: System diagram of the TDT tracking

3. Probabilistic Models

Classical IR measures compare queries with stories by using somewhat ad hoc measures that are related to how many times each query word occurs in a document. We propose to use probabilistic measures wherever possible so that we can formally express what quantities we are computing.

We use two different fundamental models for comparing a story to a group of stories on a topic:

1. The group is the model, and the words in the story were generated according to the word distribution of the group (e.g., the BBN topic spotting model).
2. The story is the model, and the words in the group were generated according to the word distribution of the story (e.g., the BBN IR metric).

In the first case, we are trying to calculate $p(T|S)$ where S is the story and T represents the group of stories on a topic. In case two, we calculate $p(SisR|T)$, which is the probability that S is relevant given the topic model.

We enhance this model further by allowing the words in a story to be generated from two word distributions: the topic-specific distribution and the general English distribution. This model is depicted in figure 2.

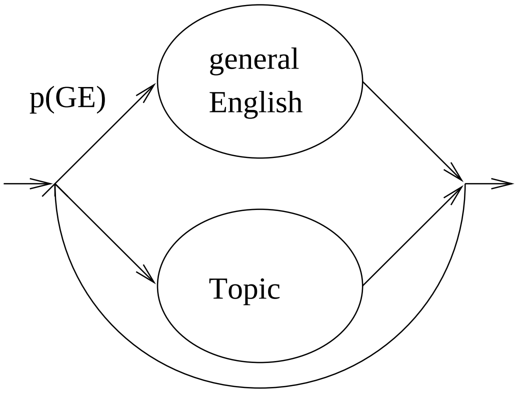

Figure 2: Two-state model for generating words in a story about a topic

3.1. Topic Spotting

The BBN topic spotting metric (TS) [6] is one method of estimating $p(T|S)$; in other words, we want to compute the posterior probability that story S comes from the distribution of topic T. By Bayes' Rule:

$$p(T|S) = p(T) \cdot \frac{p(S|T)}{p(S)}$$

where $p(T)$ is the *a priori* probability that any new story will be on topic T. Furthermore, by making an assumption that the story words are conditionally independent, we get:

$$p(T|S) \approx p(T) \cdot \prod_n \frac{p(s_n|T)}{p(s_n)}$$

where s_n corresponds to the individual words in the story, and $p(s_n|T)$ is the probability that a word in a story on topic T would be s_n.

We model $p(s_n|T)$ with a two-state mixture model, where one state is a distribution of the words in all of the stories in the group, and the other state is a distribution from the whole corpus. That is, we have a generative model for the words in the new story.

To calculate the distributions of the states, we use the Maximum Likelihood (ML) ratio estimate, which is the number of occurrences of s_n among the topic stories divided by the number of words in topic stories. This estimate can be corrected for two main weaknesses:

1. The "stop words" (e.g., the, to) dominate the score. These words can simply be eliminated.

2. The unobserved words for the topic have zero probability. Therefore, the model can be smoothed with a "back-off" to the General English model:

$$p'(s_n|T) = \frac{p(s_n|T)}{p(s_n)}$$

The estimates for the general English distribution and topic distributions can be refined using the Expectation-Maximization (EM) algorithm. This process allows new words to be added to the distributions and emphasizes topic-specific words. Therefore, the EM algorithm assigns higher probabilities to words that are more likely to be in the topic.

3.2. Information Retrieval

The BBN IR metric [8] [9] looks at the problem in exactly the opposite way. Given a query Q, we want to know the probability that any new story S is relevant to the query. But in this case, we assume that the query was generated by a model estimated from the story.

$$p(SisR|Q) = p(SisR) \cdot \frac{p(Q|S)}{p(Q)}$$

Dropping $p(Q)$ and assuming independence of words in the query, we have:

$$p(SisR|Q) \approx p(SisR) \cdot \prod_n p(q_n|S)$$

Again, we use a two-state model, where one state is a unigram distribution estimated from the story S, and the other is the unigram distribution from the whole corpus.

For the tracking problem, we use all of the stories given to be on the topic as the query. Thus, the query is a very long sequence of words – typically much longer than the new story.

3.3. Relevance Feedback

The Relevance Feedback (RF) measure is similar to the IR measure, except we do not use all of the words in the topic stories. Instead, we only use those words that are common to at least two of the on-topic stories. Each common word is used only once, but the "back-off" weight from the story state to the general English state is estimated as a function of the number of topic stories that have that query word.

4. Important Techniques

There are lots of things we tried during the past year. Most of them don't work, some help little. Only a few really improve system performance by a great deal. These are score normalization, unsupervised adaptation, and model combination. Score normalization has been a difficult task in the IR area; we find a nice solution that is simple, automatic and effective. In addition, we have found that using a time-decay prior also helps. In the next few subsections, we will illustrate the key techniques in detail.

4.1. Score Normalization

Single threshold for all topics requires score normalization across topics for optimum system performance. Normalization is a critical component of tracking systems. We collect statistics on the scores of the known off-topic and on-topic stories, then normalize the test scores based on these statistics for each topic.

There are hundreds to thousands of off-topic stories for each topic. An on-topic story can also be interpreted as not off-topic. As long as we have a reliable statistical distribution for the off-topic stories, there are existing statistical methods to identify on-topic stories based on their scores in the distribution. We use statistical hypothesis test method. A story will be marked as on-topic if its normalized score is extremely high relative to the off-topic distribution. With a threshold based on the statistical distribution of the off-topic stories, it is also relatively easy to control the false alarm error rate. Once normalized, scores from different systems are more likely to be on the same basis and directly comparable. So it makes more sense to combine systems that all produce normalized scores. Normalized scores can potentially improve the logistic regression based estimation of system combination coefficients.

The scores are normalized by the robust estimate of mean and standard deviation obtained from the off-topic stories by the following formula:

$$score' = f_{on-topic} \times \frac{score - \mu_{of-topic}}{\sigma_{of-topic}}$$

We set the scalar function $f_{on-topic}$ to be 1 for all topics, with the assumption that the average score of on-topic stories within any topic will be similar and comparable after the normalization. We will discuss later in section 5, that a better non-constant scalar function $f_{on-topic}$ can improve system performance significantly because of the variation in topic definition and its scope.

This normalization is automatic, simple and effective. The improvement after score normalization is especially significant for the IR system, as illustrated in figure 3.

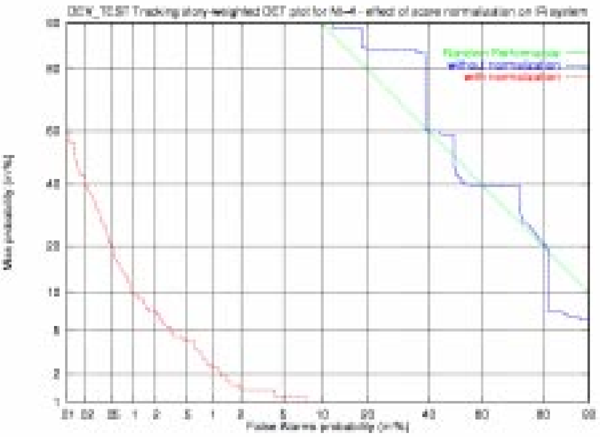

Figure 3: Effect of automatic score normalization on dev-test NWT+ASR, trained on four on-topic stories

4.2. Unsupervised Adaptation

More on-topic stories usually lead to better probabilistic models. Besides that, causal adaptation also make it possible to have the models to follow the target topics, which is an essential goal of TDT tracking task. So it is natural to always add more on-topic stories in training, given that they are, or at least very likely, on the topic. Causal unsupervised adaptation has been implemented for two of the individual systems (TS and RF). In brief, the casual unsupervised adaptation algorithm looks for a test story with very high score, adds it as an on-topic story and re-train the system before working on the next test story. Figure 4 shows the improvement on TS system, contributed solely by the unsupervised adaptation.

At the false alarm error of 0.1%, unsupervised adaptation cuts the miss error by more than 10% absolutely. Similar improvement is also observed for the RF system.

4.3. Model Combination

Different systems focus on different features of the stories. Thus, it seems reasonable to combine the probability scores from many tracking systems. From the statistical point of view, the combined system is also more likely to produce a better estimate with smaller variance.

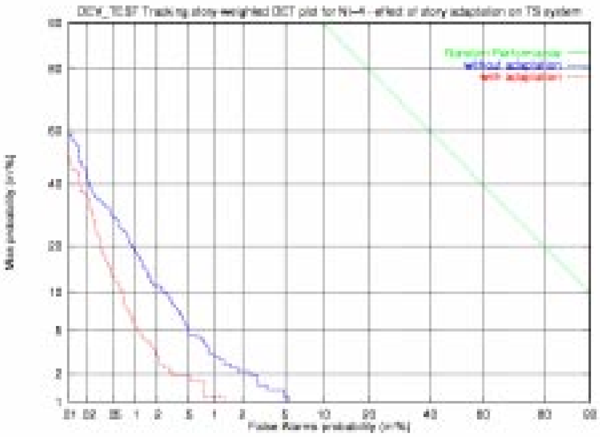

Figure 4: Effect of unsupervised adaptation for TS measure on dev-test NWT+ASR, trained on four on-topic stories

We use a linear combination of the log scores from the above three systems and the time decay to form the BBN tracking system. Our experiments show a significant reduction of both miss and false alarm rates with the combined system. The combination coefficients can be estimated by logistic regression modeling on all the stories from the TDT-2 development data set. Since the data is mostly off-topic, it is important to focus more on the on-topic stories in the logistic regression modeling. We did cross-validation experiments, and found the coefficient estimates are quite robust and reliable.

Once the system is combined as in figure 5, we see a significant improvement over each individual systems (TS, IR and RF). At 0.1% false alarm error, the best individual system has 6.5% miss error and the combined system only has 3%. So in the area where TDT tracking task has most interests, the miss error rate is cut by more than half with system combination of 3 individual systems.

4.4. Decay Prior

Different systems focus on different features of the stories. We tried to utilize time information that is available for each story. It seems reasonable to combine the probability scores from many tracking systems with a time-decayed prior probability score. This reflects that a test story is less likely to be on-topic as its age in-

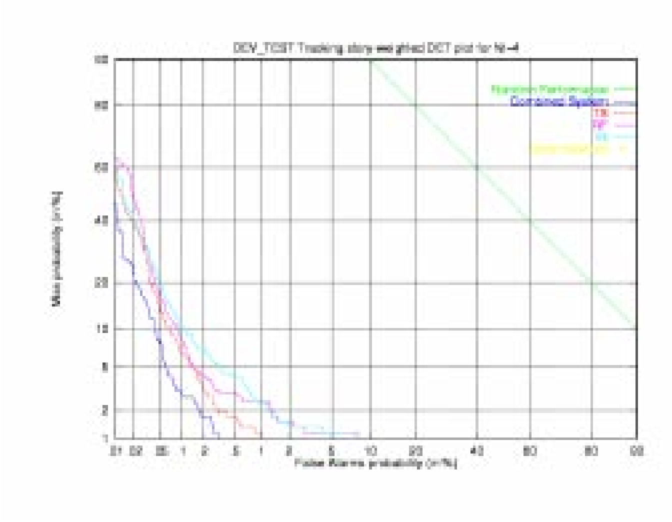

Figure 5: Combined system on dev-test NWT+ASR, trained on four on-topic stories

creases.

5. BBN Tracking System Performance

We developed and tested the BBN tracking system on the dev-test TDT-2 data. At 0.1% false alarm error, the system achieves a performance of 1.5% miss rate (NWT+ASR) and 3.0% miss rate (NWT+CCAP). This indicates that the automatic tracking system can do as well as the experienced human annotators who make at least 6% miss errors.

The same system was used in the 1998 DARPA evaluation on the TDT-2 data collected during May and June. The Ctrack score, a compromise over miss errors and false alarm errors $(0.002P(miss) + 0.098P(fa))$, is the official error measure to evaluate the performance of all the tracking systems for both story-weighted and topic-weighted. For our tracking system evaluated on NWT+ASR data, the Ctrack numbers are 0.0064 and 0.0057 respectively. These are much higher than what we have seen for the dev-test, where the Ctrack numbers are as low as 0.0018. The evaluation data set is likely to be quite different and harder than the dev-test. One of the observation is the significant higher variation of number of on-topic stories in the evaluation data. In fact, the three largest topics in the evaluation have more than 900 on-topic stories, where the total number of on-topic stories are about 600 and 1500 respectively for the dev-test and the evaluation data. Since story-weighted Ctrack measure is very biased to the large topics, the tracking algorithm needs to address the issue of significant variation in the definition of the topic and its scope.

After the evaluation, we noticed that the average score of on-topic stories are not always similar and comparable after normalization. For the evaluation data, the average score of the 4 given training on-topic stories varies from 21.1 to 54.5, and it is positively correlated to the average score of the testing on-topic stories. This motivates us to use the mean and standard deviation collected from the given on-topic stories to adjust the final score so that the average score of testing on-topic stories can be similar and comparable. Using the non-constant scalar function $f_{on-topic} = \frac{25}{\mu_{on-topic} - 0.5\sigma_{on-topic}}$ on the 1998 evaluation data, we see a significant overall improvement on the DET curve in figure 6. Ctrack decreased dramatically from 0.0064 to 0.0045 for story-weighted measure, and modestly from 0.0057 to 0.0052 for topic-weighted measure.

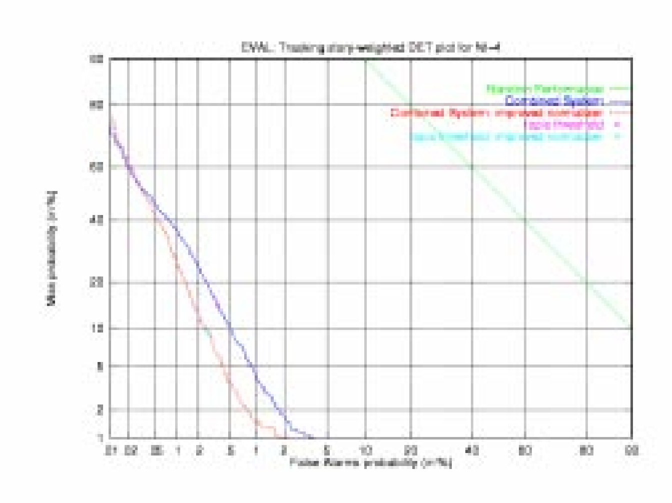

Figure 6: Effect of the non-constant scalar function in normalization, evaluation data, NWT+ASR, trained on four on-topic stories

We also compared the system performance of using automatic speech recognition output (NWT+ASR) versus the human transcribed closed caption text (NWT+CCAP). For both dev-test and evaluation data, we see that NWT+CCAP has little performance edge on NWT+ASR in the DET curves. This overall difference between NWT+ASR and NWT+CCAP may not even be statistically significant. However, at the lower end of the DET curves where low miss rate and high false alarm rate can be achieved, the tracking system does work noticeably better on NWT+CCAP than on NWT+ASR.

6. CONCLUSION

We developed a probability based tracking system utilizing advanced technologies such as automatic score normalization, unsupervised adaptation and model combination. Our tracking system was ranked the best in the 1998 TDT-2 evaluation. Tested on the TDT-2 data, this tracking system achieves a performance comparable to experienced human annotators. We introduced new approaches for score normalization and system combination. Our statistics based score normalization is simple, automatic and effective. We believe this statistics based score normalization algorithm can be adapted and generalized to other IR applications where normalization is always a hard problem. And finally, model combination is statistically sound and can significantly improve performance over individual systems as we have shown from our experimental results.

7. ACKNOWLEDGMENTS

This work was supported by the Defense Advanced Research Projects Agency and monitored by Ft. Huachuca under contract No. DABT63-94-C-0063 and by the Defense Advanced Research Projects Agency and monitored by NRaD under contract No. N66001-97-D-8501. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

References

1. "The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan," Version 3.7. August 3, 1998.
2. Charles Wayne, "Topic Detection & Tracking (TDT) Overview & Perspective", DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, February, 1998.
3. George Doddington, "The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan", DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, February, 1998.
4. L. Gillick Y. Ito, L. Manganaro, M. Newman, F. Scattone, S. Wegmann, J. Yamron, P. Zhan, "Dragon Systems' Automatic Transcription of New TDT Corpus", DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, February, 1998.
5. J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. "Topic Detection and Tracking Pilot Study Final Report." *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.* February, 1998.
6. R. Schwartz, T. Imai, L. Nguyen, and J. Makhoul. "A Maximum Likelihood Model for Topic Classification of Broadcast News." *Eurospeech '97*, Rhodes, Greece. September, 1997.
7. G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, NY. 1983.
8. T. Leek, D. Miller, and R. Schwartz, "Labrador: A Hidden Markov Model Information Retrival System", submitted to SIGIR-99.
9. F. Walls, H. Jin, S. Sista, R. Schwartz, "Probabilistic models for topic detection and tracking", ICASSP, March, 1999.